**Exercise – 3**

**Clean Data in Excel:**

**1) Get Rid of Extra Spaces:**

Extra Spaces are difficult to spot & correct. Multiple spaces may be easy, but trailing spaces are pretty tough. Trailing spaces are blank spaces at the end of the statement or word which are not followed by any other character.

Here’s an easy way to spot & eliminate such errors.

**Syntax:** TRIM(text)

**Steps:**

* Consider data with four cells with different spacing errors.
* Now select a column & type “**TRIM(**“
* Now select the cell you want to correct (in matters of spaces).
* The cell will be corrected. If there are other erroneous cells sequentially aligned, drag the fixed cell till the point, you want to check & correct.

This easy step can save you time!

**2) Select & Treat all blank cells:**

Blank cells are troublesome because they often create errors while creating reports. And, people usually want to replace such cells with 0, Not Available or something like that. But replacing each cell manually on a large data table would take hours. Luckily, there’s an easy way to tackle this problem.

**Steps:**

* Select the entire Data (you want to treat)
* Press **F5** (on keyboard)
* A dialogue box will appear > Select “**Special**”
* Select “**Blanks**” & click “**OK**”
* Now, all blank cells will be highlighted in pale grey color, out of which one cell would be white with a different border. That’s the active cell, type the statement you want to replace in blank cells.
* Hit “**Ctrl+Enter**”

**NOTE:**

At the last step, if “**Enter**” only is pressed, then the value will be inserted only in the active cell. So remember to press “**Ctrl+Enter**.”

**3) Convert Numbers Stored as Text into Numbers:**

When we import data from files, other sources, databases, text, etc. During transit, data might get affected. Also, some have a habit of using an apostrophe before numerical values, which is considered as text in Excel. Such minor data conversion can drastically affect calculations.

Suppose there are three values “70, ’70, 80”. When we compare 70 and 80 (70<80), the result is “TRUE.” But when we compare “apostrophe 70 & 80” (‘70<80), the problem starts. Here the result will be FALSE as the text will be rated higher than any number. To eliminate such errors, here’s a trick.

**Steps:**

* Select any blank cell & type**1**
* Select that cell & hit “**Ctrl+C**”
* Now select your data set & go to **Paste > Paste Special**
* In Paste Special, select “**Multiply**” option in the “**Operation**” category
* Click “**OK**”

Here it multiples every single value to “1”. And anything multiplied by 1 is the same number. But this trick also takes care of the apostrophe numerical.

**4) Remove Duplicates:**

Elimination of duplicate data is necessary for the creation of unique data & less usage of storage. In duplication, you can either highlight it or delete it.

**A) Highlight Duplicates:**

* Select the data & go to **Home > Conditional Formatting > Highlight Cell** **Rules > Duplicate Values**
* A dialogue box will appear (Duplicate Values), Select **Duplicate & formatting color**
* Press **OK**
* All duplicate values will be highlighted!

**B) Delete Duplicates:**

* Select the data & go to **DATA > Remove Duplicates**
* A dialogue box will appear (Remove Duplicates), tick columns whose duplicates need to be found.
* Remember to have a click on “**My data has headers**” (if your Data has headers) or else column heads will be considered as data & duplication search will be applied on it too.
* Click OK!

Duplicate values will be removed! Suppose you select 4 of 4 columns. Then that four columns rows should also match or else; they won’t be considered duplicate.

**5) Highlight Errors:**

While creating reports or dashboards, you might face a few arithmetical errors (like divisional errors). Such errors are easy to spot if the Data is small. But for big data, it’s complicated. So to get rid of such mistakes, you can go for two ways: Conditional Formatting or Go to Special.

**A) Using Conditional Formatting:**

* Select the Data
* Go to **Home > Conditional Formatting > New Rule**
* Within New Rule, Select “**Format only cells that contain**.”
* In Rules, Select “**Errors” & Click on “Format**”
* Select any color & click **OK**
* Hit the final “**OK**” button

All the cells with errors are highlighted & now are easy to spot.

**B) Go to Special:**

* Select the Data
* Press **F5**
* Click on “Special”
* A dialogue box appears (Go to Special), Select Formulas
* Now you get four options in Formulas, deselect all options except “Errors”
* Click OK! Now all errors are selected, you can delete them manually or replace a statement.
* If you wish to replace, then type the statement at active cell & hit “**CTRL+ENTER**.”

**6) Change Text to Lower/Upper/Proper Case:**

While importing data, we often find names in irregular forms like a lower, upper case, or sometimes mixed. Such errors are not easy to eliminate manually. Here’s a fingertip trick to bring back the consistency.

* LOWER(text)
* UPPER(text)
* PROPER(text)

**Steps:**

* Just type the formula you want to use, suppose “**LOWER(**“ and select the cell whose case needs to be changed.
* Hit “**CTRL+ENTER**.”
* The case has been changed & consistent
* Drag down to do the same for other cells.
* Similarly for **UPPER()** &**PROPER()**

**7) Parse Data Using Text to Column:**

Sometimes the received Data has texts filled in one cell, only separated by punctuations. Usually, the addresses are cramped in one cell separated by a comma. To distinguish values in separate cells, we can use “Text to Column.”

**Steps:**

* Select the Data
* Go to **Data> Text to Column**
* A dialogue box will appear (Convert Text to Columns Wizard – Step 1 of 3), select Delimited or Fixed Width as per your convenience.
* Delimited is to be selected if the width isn’t fixed, click “**NEXT**”
* In Delimiters tick the option which separates your text in the cell. Suppose “Norwich Cathedral, Norwich, UK,” here three values are separated by commas. So we will select “Comma” for this example. And, deselect rest options.
* View the preview & click on “**NEXT**”
* Select Column Data Format & destination cell address

Click “**FINISH**”

**8) Spell Check:**

Spelling mistakes are common in text files & PowerPoint. However, MS points out such errors by underlining it with colorful dashes. And, MS Excel doesn’t have such feature. But you can use it below steps:

* Select the Data
* Press “**F7**”
* A dialogue box appears, which shows you the possible wrong word & it’s the possible correct spelling. Click on “**Change**,” if you agree with the suggestion.
* Check & change till it says “**Spell check complete. You’re good to go!**

**9) Delete all Formatting:**

Suppose you want to clear all the formats, including highlights & borders. You can do this by selecting the data & go to HOME > Clear (in editing group) > Clear Formats. It will clear the formats & you get standard content without highlights or borders. Similarly, you can clear Content, Comments, Hyperlink, or entire data (using Clear All).

**10) Use Find & Replace to Clean Data in Excel:**

**A) Changing Cell References:**

* Press “**CTRL+H**” to open “**Find and Replace**”
* Now in **Replace > “Find What**” (change the reference range too) “**Replace With**”
* Suppose Find What: $B to Replace With: $C
* Click on “**Replace All**”
* Similarly finding & replacing using reference range we can clean the Data

**B) Find & Change Specific Format:**

* Press “**CTRL+H**”
* Select “**Options**”
* Now go to “**Format” of “Find What**.” Here you can specify the format or choose a format from the cell. Suppose you select a format.
* Now it will show you the preview for “**Find What**.”
* Click on “**Format**” of “**Replace With**.” Suppose we go for “Format…”
* Now select format, example: Number, Alignment, Font, Border, Fill, Protection.
* Suppose we select Color then select any color to fill the column header cell.
* Click on Replace All
* Instantly the format has been changed!

**C) Removal of Line Breaks:**

Suppose we have a data where it is separated by line breaks (same cell but different rows). To remove these line breaks, follow the below steps:

* Press “**CTRL+H**”
* Find and Replace dialogue box will appear, press “**CTRL+J**”
* Go to the replace with box & type a single space
* Click Replace All
* All rows will be managed in one row within the same cell!

**D) Removal of Parenthesis:**

* Select the Data
* Press “**CTRL+H**”
* Type (\*) in “Find What” (This will consider all characters within parenthesis)
* Leave the Replace With column empty & click Replace
* Parenthesis characters are removed!

**EXAMPLES**

Initially, you need to clean the data. Data cleaning includes removing unwanted characters from text. Next, you need to structure the data in the form you require for further analysis. You can do the same by −

* Finding required text patterns with the text functions.
* Extracting data values from text.
* Formatting data with text functions.
* Executing data operations with the text functions.

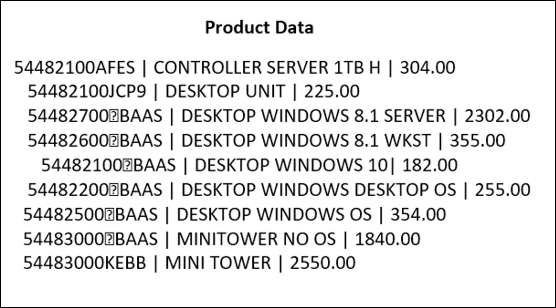
**Removing Unwanted Characters from Text:**

When you import data from another application, it can have nonprintable characters and/or excess spaces. The excess spaces can be −

* leading spaces, and/or
* extra spaces between words.

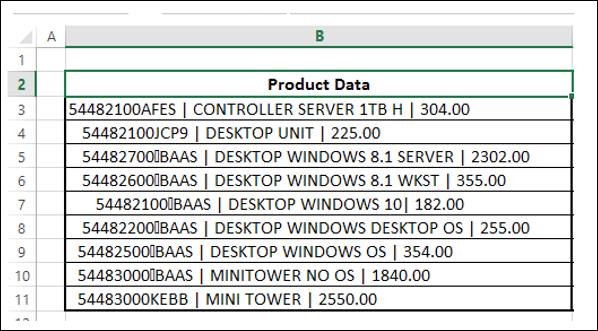
If you sort or analyze such data, you will get erroneous results.

Consider the following example −



This is the raw data that you have obtained on product information containing the Product ID, Product description and the price. The character “|” separates the field in each row.

When you import this data into Excel worksheet, it looks as follows −



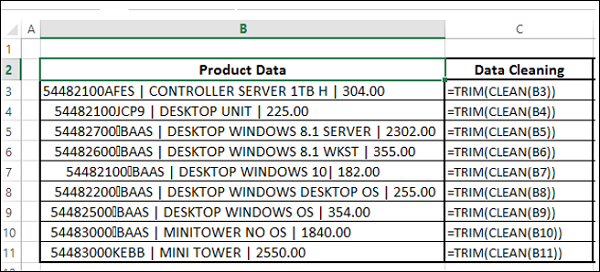
As you observe, the entire data is in a single column. You need to structure this data to perform data analysis. However, initially you need to clean the data.

You need to remove any nonprintable characters and excess spaces that might be present in the data. You can use the CLEAN function and TRIM function for this purpose.

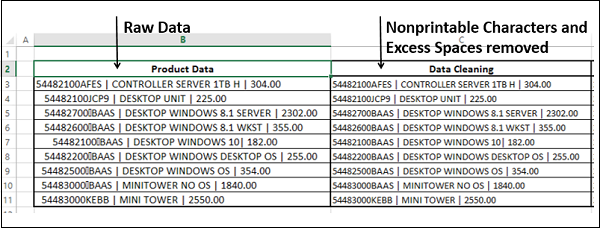
|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **CLEAN**  Removes all nonprintable characters from text |
| 2. | **TRIM**  Removes spaces from text |

* Select the Cells C3 – C11.
* Type =TRIM (CLEAN (B3)) and then press CTRL + Enter.

The formula is filled in the cells C3 – C11.



The result will be as shown below −



**Finding required Text Patterns with the Text Functions:**

To structure your data, you might have to do certain Text Pattern matching based on which you can extract the Data Values. Some of the Text Functions that are useful for this purpose are −

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **EXACT**  Checks to see if two text values are identical |
| 2. | **FIND**  Finds one text value within another (case-sensitive) |
| 3. | **SEARCH**  Finds one text value within another (not case-sensitive) |

**Extracting Data Values from Text:**

You need to extract the required data from text in order to structure the same. In the above example, say, you need to place the data in three columns – ProductID, Product\_Description and Price.

You can extract data in one of the following ways −

* Extracting Data Values with Convert Text to Columns Wizard
* Extracting Data Values with Text Functions
* Extracting Data Values with Flash Fill

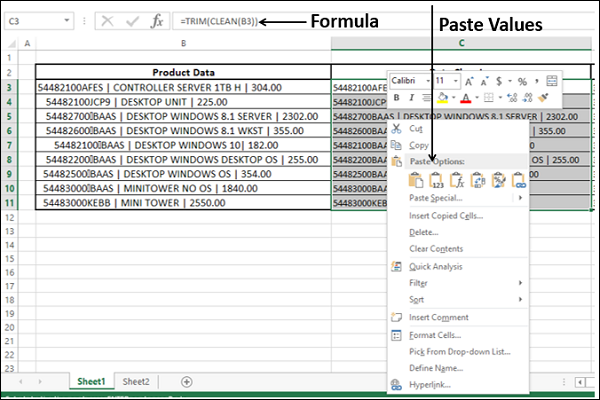
**Extracting Data Values with Convert Text to Columns Wizard:**

You can use the **Convert Text to Columns Wizard** to extract Data Values into Excel columns if your fields are −

* Delimited by a character, or
* Aligned in columns with spaces between each field.

In the above example, the fields are delimited by the character “|”. Hence, you can use the **Convert Text to Columns** wizard.

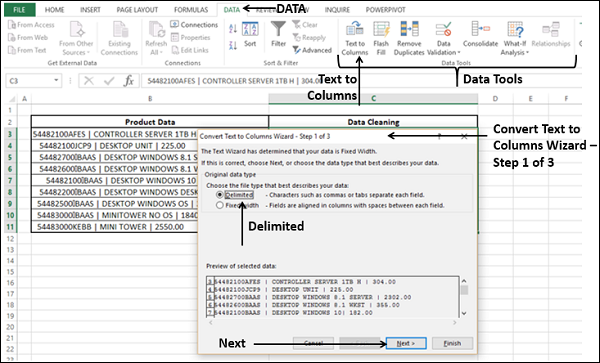
* Select the data.
* Copy and paste values in the same place. Otherwise, **Convert Text to Columns** takes the functions rather than the data itself as the input.



* Select the data.
* Click on **Text to Columns** in the **Data Tools** group under **Data** Tab on the Ribbon.

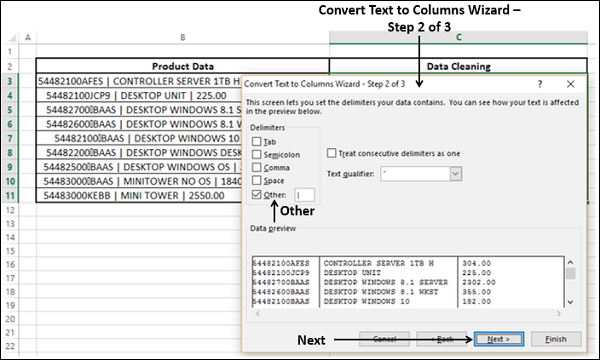
**Step 1** − Convert Text to Columns Wizard - Step 1 of 3 appears.

* Select Delimited.
* Click Next.



**Step 2** − Convert Text to Columns Wizard - Step 2 of 3 appears.

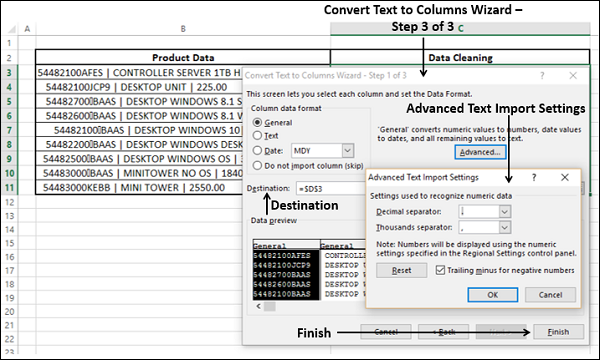
* Under **Delimiters**, select **Other**.
* In the box next to **Other**, type the character **|**
* Click **Next**.



**Step 3** − Convert Text to Columns Wizard - Step 3 of 3 appears.

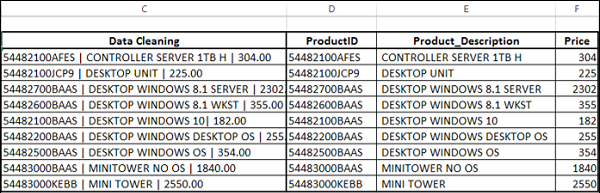
In this screen, you can select each column of your data in the wizard and set the format for that column.

* For **Destination**, select the cell D3.
* You can click **Advanced**, and set **Decimal Separator** and **Thousands Separator** in the **Advanced Text Import Settings** dialog box that appears.
* Click **Finish**.



Your data, which is converted to columns appears in the three Columns – D, E and F.

* Name the Column headers as ProductID, Product\_Description and Price.



**Extracting Data Values with Text Functions:**

Suppose the fields in your data neither are delimited by a character nor are aligned in columns with spaces between each field, you can use text functions to extract data values. Even in the case the fields are delimited, you can still use text functions to extract data.

Some of the text functions that are useful for this purpose are –

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **LEFT**  Returns the leftmost characters from a text value |
| 2. | **RIGHT**  Returns the rightmost characters from a text value |
| 3. | **MID**  Returns a specific number of characters from a text string starting at the position you specify |
| 4. | **LEN**  Returns the number of characters in a text string |

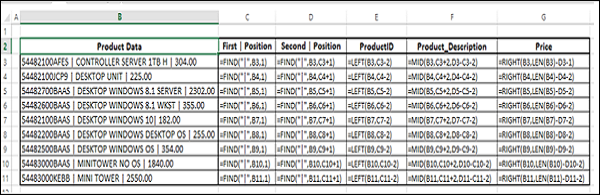
You can also combine two or more of these text functions as per the data you have at hand, to extract the required data values. For example, using a combination of LEFT, RIGHT and VALUE functions or using a combination of FIND, LEFT, LEN and MID functions.

In the above example,

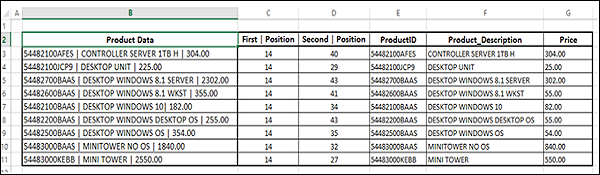
* All the characters left to the first | give the name ProductID.
* All the characters right to the second | give the name Price.
* All the characters that lie between the first | and second | give the name Product\_Description.
* Each | has a space before and after.

Observing this information, you can extract the data values with the following steps −

* Find the Position of First | - **First | Position**
  + You can use FIND function
* Find the Position of Second | - **Second | Position**
  + You can use FIND function again
* Beginning to (**First | Position** – 2) Characters of the Text give ProductID
  + You can use LEFT Function
* (**First | Position** + 2) to (**Second | Position** - 2) Characters of the Text give Product\_Description
  + You can use MID Function
* (**Second | Position** + 2) to End Characters of the Text give Price
  + You can use RIGHT Function



The result will be as shown below −

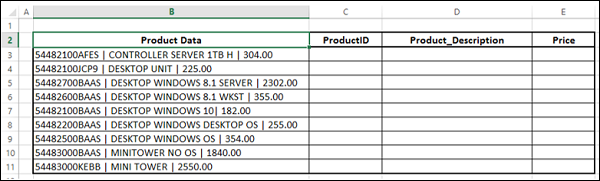


You can observe that the values in the price column are text values. To perform calculations on these values, you have to format the corresponding cells. You can look at the section given below to understand formatting text.

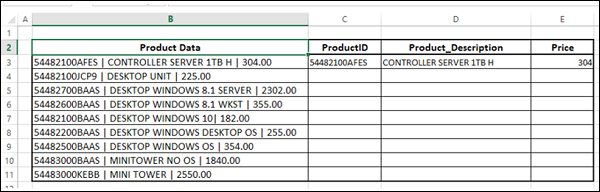
Extracting Data Values with Flash Fill:

Using Excel **Flash Fill** is another way to extract data values from text. However, this works only when Excel is able to find a pattern in the data.

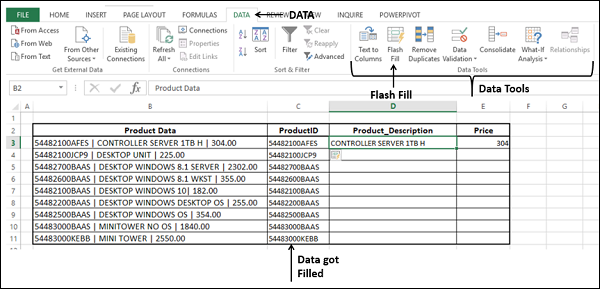
**Step 1** − Create three columns for ProductID, Product\_Description and Price next to the data.



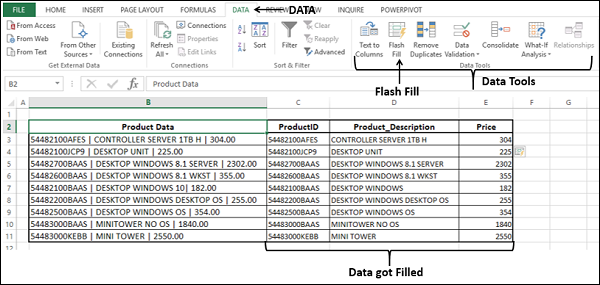
**Step 2** − Copy and paste the values for C3, D3 and E3 from B3.



**Step 3** − Select cell C3 and click **Flash Fill** in the **Data Tools** group on the **Data** tab. All the values for ProductID get filled.



**Step 4** − Repeat the above given steps for Product\_Description and Price. The data is filled.



**Formatting Data with Text Functions:**

Excel has several built-in text functions that you can use for formatting data containing text. These include −

**Functions that format the Text as per your need** −

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **LOWER**  Converts text to lowercase |

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **UPPER**  Converts text to uppercase |
| 2. | **PROPER**  Capitalizes the first letter in each word of a text value |

**Functions that convert and/or format the Numbers as Text** −

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **DOLLAR**  Converts a number to text, using the $ (dollar) currency format |
| 2. | **FIXED**  Formats a number as text with a fixed number of decimals |
| 3. | **TEXT**  Formats a number and converts it to text |

**Functions that convert the Text to Numbers** −

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **VALUE**  Converts a text argument to a number |

**Executing Data Operations with the Text Functions:**

You might have to perform certain Text Operations on your Data. For example, if Login-IDs for the Employees are changed to a New Format in an Organization, based on the Format Change, Text Replacements might have to be done.

Following Text Functions help you in performing Text Operations on your data containing Text −

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **REPLACE**  Replaces characters within text |
| 2. | **SUBSTITUTE**  Substitutes new text for old text in a text string |
| 3. | **CONCATENATE**  Joins several text items into one text item |
| 4. | **CONCAT**  Combines the text from multiple ranges and/or strings, but it does not provide the delimiter or IgnoreEmpty arguments. |
| 5. | **TEXTJOIN**  Combines the text from multiple ranges and/or strings, and includes a delimiter you specify between each text value that will be combined. If the delimiter is an empty text string, this function will effectively concatenate the ranges. |
| 6. | **REPT**  Repeats text a given number of times |

**Exercise -4**

The data that you obtain from different sources might contain date values. In this chapter, you will understand how to prepare your data that contains data values for analysis.

You will learn about −

* Date Formats
  + Date in Serial Format
  + Date in different Month-Day-Year Formats
* Converting Dates in Serial Format to Month-Day-Year Format
* Converting Dates in Month-Day-Year Format to Serial Format
* Obtaining Today's Date
* Finding a Workday after specified Days
* Customizing the Definition of a Weekend
* Number of Workdays between two given Dates
* Extracting Year, Month, Day from Date
* Extracting Day of the Week from Date
* Obtaining Date from Year, Month and Day
* Calculating Number of Years, Months and Days between two Dates

**Date Formats:**

Excel supports **Date** values in two ways −

* Serial Format
* In different Year-Month-Day Formats

You can convert −

* A **Date** in Serial Format to a **Date** in Year-Month-Day Format
* A **Date** in Year-Month-Day Format to a **Date** in Serial Format

**Date in Serial Format:**

A **Date** in serial format is a positive integer that represents the number of days between the given date and January 1, 1900. Both the current **Date** and January 1, 1900 are included in the count. For example, 42354 is a **Date** that represents 12/16/2015.

**Date in Month-Day-Year Formats:**

Excel supports different **Date** Formats based on the **Locale** (Location) you choose. Hence, you need to first determine the compatibility of your **Date** formats and the Data Analysis at hand. Note that certain **Date** formats are prefixed with \*(asterisk) −

* **Date** formats that begin with \*(asterisk) respond to changes in regional date and time settings that are specified for the operating system
* **Date** formats without an \*(asterisk) are not affected by operating system settings

For understanding purpose, you can assume United States as the Locale. You find the following **Date** formats to choose for the **Date** - 8th June, 2016 −

* \*6/8/2016 (affected by operating system settings)
* \*Wednesday, June 8, 2016 (affected by operating system settings)
* 6/8
* 6/8/16
* 06/08/16
* 8-Jun
* 8-Jun-16
* 08-Jun-16
* Jun-16
* June-16
* J
* J-16
* 6/8/2016
* 8-Jun-2016

If you enter only two digits to represent a year and if −

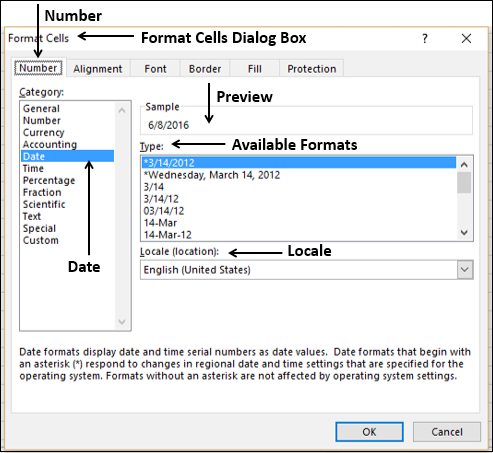
* The digits are 30 or higher, Excel assumes the digits represent years in the twentieth century.
* The digits are lower than 30, Excel assumes the digits represent years in the twenty-first century.

For example, 1/1/29 is treated as January 1, 2029 and 1/1/30 is treated as January 1, 1930.

**Converting Dates in Serial Format to Month-Day-Year Format:**

To convert dates from serial format to Month-Day-Year format, follow the steps given below −

* Click the **Number** tab in the **Format Cells** dialog box.
* Click **Date** under **Category**.
* Select **Locale**. The available **Date** formats will be displayed as a list under **Type**.
* Click on a **Format** under **Type** to look at the preview in the box adjacent to **Sample**.



After choosing the Format, click **OK.**

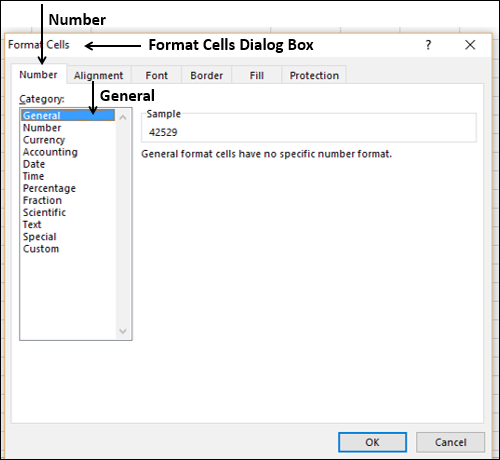
**Converting Dates in Month-Day-Year Format to Serial Format:**

You can convert dates in Month-Day-Year format to Serial format in two ways −

* Using **Format Cells** dialog box
* Using Excel **DATEVALUE** function

Using Format Cells dialog box

* Click the **Number** tab in the **Format Cells** dialog box.
* Click **General** under **Category**.



**Using Excel DATEVALUE Function:**

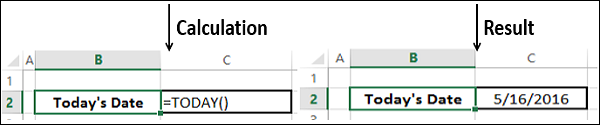
You can use Excel **DATEVALUE** function to convert a **Date** to **Serial Number** format. You need to enclose the **Date** argument in “”. For example,

=DATEVALUE ("6/8/2016") results in 42529

Obtaining Today's Date

If you need to perform calculations based on today’s date, simply use the Excel function TODAY (). The result reflects the date when it is used.

The following screenshot of TODAY () function usage has been taken on 16th May, 2016 −



**Finding a Workday after Specified Days:**

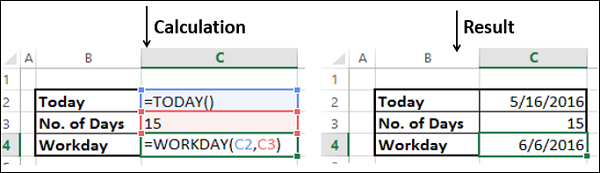
You might have to perform certain calculations based on your workdays.

Workdays exclude weekend days and any holidays. This means if you can define your weekend and holidays, whatever calculations you do will be based on workdays. For example, you can calculate invoice due dates, expected delivery times, the next meeting date, etc.

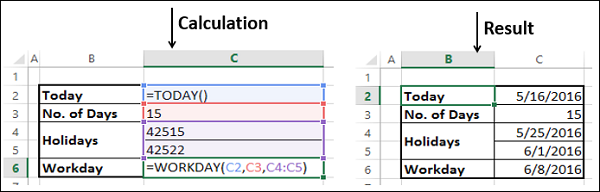
You can use Excel **WORKDAY** and **WORKDAY.INTL** functions for such operations.

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **WORKDAY**  Returns the serial number of the date before or after a specified number of workdays |
| 2. | **WORKDAY.INTL**  Returns the serial number of the date before or after a specified number of workdays using parameters to indicate which and how many days are weekend days |

For example, you can specify the 15th working day from today (the screenshot below is taken on 16th May 2016) using the Functions TODAY and WORKDAY.



Suppose 25th May 2016 and 1st June 2016 are holidays. Then, your calculation will be as follows −

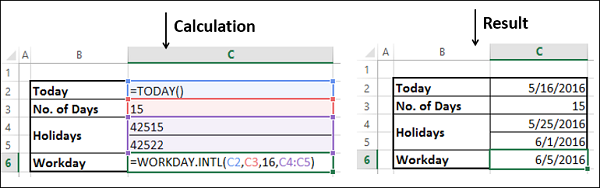


**Customizing the Definition of a Weekend:**

By default, weekend is Saturday and Sunday, i.e. two days. You can also optionally define your weekend with the **WORKDAY.INTL** function. You can specify your own weekend by a weekend-number that corresponds to the weekend days as given in the table below. You need not remember these numbers, because when you start typing the function, you get a list of numbers and the weekend days in the drop-down list.

|  |  |
| --- | --- |
| **Weekend Days** | **Weekend-number** |
| Saturday, Sunday | 1 or omitted |
| Sunday, Monday | 2 |
| Monday, Tuesday | 3 |
| Tuesday, Wednesday | 4 |
| Wednesday, Thursday | 5 |
| Thursday, Friday | 6 |
| Friday, Saturday | 7 |
| Sunday only | 11 |
| Monday only | 12 |
| Tuesday only | 13 |
| Wednesday only | 14 |
| Thursday only | 15 |
| Friday only | 16 |
| Saturday only | 17 |

Suppose, if weekend is Friday only, you need to use the number 16 in the WORKDAY.INTL function.



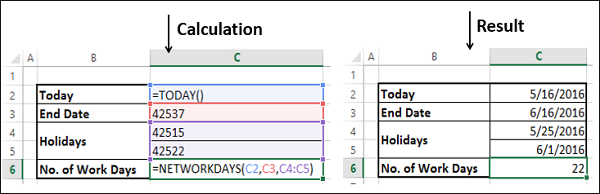
**Number of Workdays between two given Dates:**

There might be a requirement to calculate the number of workdays between two dates, for example, in the case of calculating payment to a contract employee who is paid on per day basis.

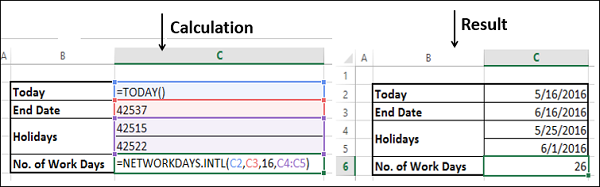
You can find the number of workdays between two dates with the Excel functions **NETWORKDAYS** and **NETWORKDAYS.INTL**. Just as in the case of WORKDAYS and WORKDAYS.INTL, NETWORKDAYS and NETWORKDAYS.INTL allow you to specify holidays and with NETWORKDAYS.INTL you can additionally specify the weekend.

|  |  |
| --- | --- |
| **S.No.** | **Function & Description** |
| 1. | **NETWORKDAYS**  Returns the number of whole workdays between two dates |
| 2. | **NETWORKDAYS.INTL**  Returns the number of whole workdays between two dates using parameters to indicate which and how many days are weekend days |

You can calculate the number of workdays between today and another date with the functions TODAY and NETWORKDAYS. In the screen shot given below, today is 16th May 2016 and end date is 16th June 2016. 25th May 2016 and 1st June 2016 are holidays.



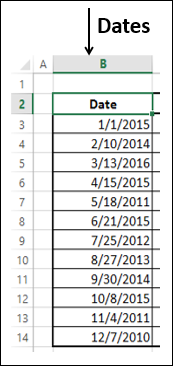
Again, the weekend is assumed to be Saturday and Sunday. You can have your own definition for weekend and calculate the number of workdays between two dates with the NETWORKDAYS.INTL function. In the screen shot given below, only Friday is defined as weekend.



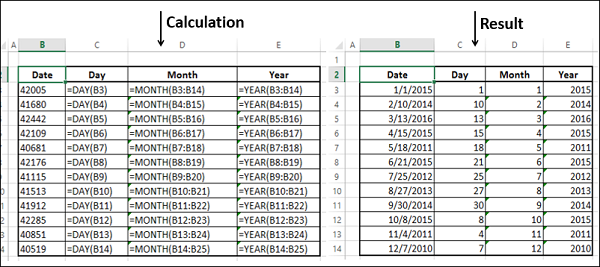
**Extracting Year, Month, Day from Date:**

You can extract from each date in a list of dates, the corresponding day, month and year using the excel functions DAY, MONTH and YEAR.

For example, consider the following dates −



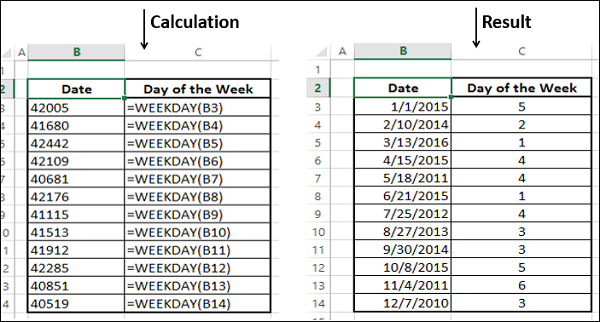
From each of these dates, you can extract day, month and year as follows −



**Extracting Day of the Week from Date:**

You can extract from each date in a list of dates, the corresponding day of the week with Excel WEEKDAY function.

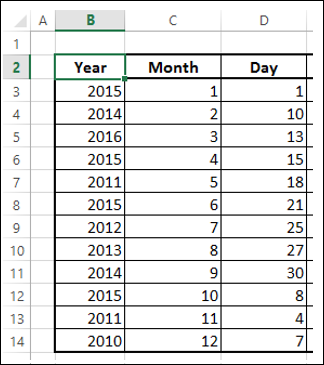
Consider the same example given above.



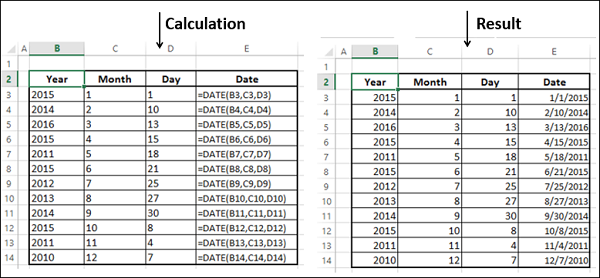
**Obtaining Date from Year, Month and Day:**

You data might have the information about Year, Month and Day separately. You need to get the date combining these three values to perform any calculation. You can use the DATE function for getting the date values.

Consider the following data −

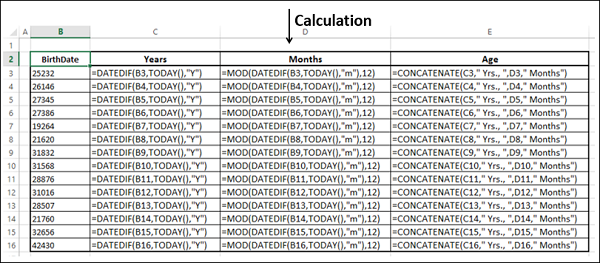


Use the DATE function to obtain DATE values.



**Calculating Years, Months and Days between two Dates:**

You might have to calculate the time lapsed from a given date. You might need this information in the form of years, months and days. A simple example would be calculating the current age of a person. It is effectively the difference between the birth date and today. You can use Excel DATEDIF, TODAY and CONCATENATE functions for this purpose.



The output is as follows −

